The first year of LHC Physics Analysis Using the Grid: Prospects from CMS

Ian Fisk
for the CMS collaboration
July 29, 2006

Introduction

CMS has had a distributed computing model from early in the experiment planning. Motivated by a variety of factors

- The large quantity of data and computing required encouraged distributed resources from a facility infrastructure point of view
- Ability to leverage resources at labs and university
  - Hardware, expertise, infrastructure

~20% of the resources are located at CERN, 40% at T1s, and 40% T2s

- The Tier-2 centers are the primary location for analysis activities
- Data selection and skimming can be performed at Tier-1 centers

There are not sufficient resources at any one center to complete the analysis tasks of the experiment

- Grid analysis will have to succeed in CMS from the opening of the experiment

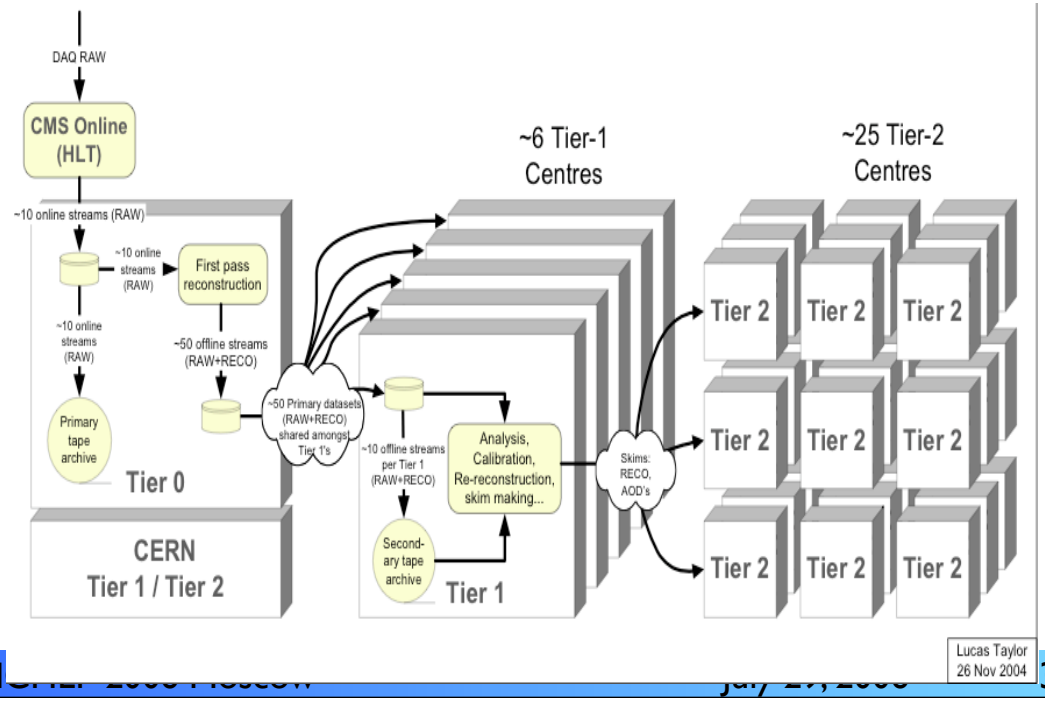
CMS Computing Model

CMS has proposed a computing model where the site activities and functionality is largely predictable

- Activities are driven by data location
  - Data is divided into streams and hosted at experiment specified sites
- Opportunistic computing is largely restricted to limited activities

A system we think we can build

- Does not prevent more dynamic computing models if functionality and capacity are available



Surviving the first years

The computing for CMS is hardest as the detector is being understood

- The analysis object data for CMS is estimated at 0.05MB
  - An entire year's data and simulation are only 300TB
    - Estimated disk for the average CMS T2 analysis center is 200TB
- Data is divided into ~10 trigger streams and ~50 offline streams
  - A physics analysis should rely on 1 trigger stream

Unfortunately, until the detector and reconstruction are completely understood the AOD is not useful for most analysis and access to the raw data will be more frequent

- The full raw data is 35 times bigger
- Given the other hadron collider experience, we can expect at least a couple of years to stabilize

The early years of the experiment will involve moving larger data objects to analysis centers

- Grid data management services need to work well at the beginning

Simplifying Constraints

CMS is applying constraints to the grid enabled sites to reduce the complexity of the computing problem

- Analysis jobs are only executed on sites that support the experiment
  - Opportunistic computing is reserved for simulated event generation
  - Sites that accept analysis jobs provide some lightweight experiment specific services and configurations
    - The data management system runs as a site service, controlling the data resident on the local site
- Data Location
  - Processing requests are sent to sites with data
    - The data that will be accessed by an analysis job is fully specified at submission
    - All data access is made over local access protocols
    - Calibration data is accessible through read-only database caches
- Software installation
  - The basic CMS software is installed on the site allowing only the user modifications to the standard distributions to be sent with the job

Managing Data for Analysis

In the CMS model analyses are performed on datasets

- Data sets may be created by the experiment as event reconstruction, analysis groups from complete data streams at Tier-1 centers, or user created skimming and selection jobs at Tier-1 or Tier-2 centers
- A dataset is registered in the CMS dataset bookkeeping service (DBS)
  - The global instance of this is a database with a defined service interface
    - Local scope instances for individual users, groups, and reconstruction tasks exist
  - The DBS knows how a group of files forms a dataset
    - CMS files are anticipated to be 5-10GB on average, currently we aim for 1-2GB
    - The DBS also maps the files into logical quantities called data blocks
  - The blocks are registered into the Dataset Location Service (DLS)
    - DLS is currently based on the grid catalog technology Local File Catalog (LFC)
    - Smallest quantity of data that the data transfer service should deal with regularly
  - Moving blocks between sites is handled by PhEDEx (Physics Experiment Data Exporter)
    - Relies on Storage Resource Manager(SRM) of the File Transfer Service (FTS)

Specifying and Submitting Applications

Once the data blocks have been located at a site the analysis jobs must be submitted

In July of 2005 CMS introduced the CMS Remote Analysis Builder (CRAB)

- CRAB was originally developed by INFN, though has grown into a global effort with contributions from the US and the UK
- A system in which a user could specify the data set desired, the application and input parameters to run, and the the number of events to process per job
  - CRAB handles the data discovery
    - Query the DBS to determine the blocks required to complete the request and then the DLS to determine the clusters that can satisfy the request
  - The job preparation
    - Tarring up the user application and parameters, while making the appropriate number of jobs for the events needed to process
  - Submitting the application
    - Submitting jobs through the appropriate grid infrastructure

CRAB Submission

A user can query the DBS to determine dataset parameters

- Current query capabilities are fairly primitive, but will improve.

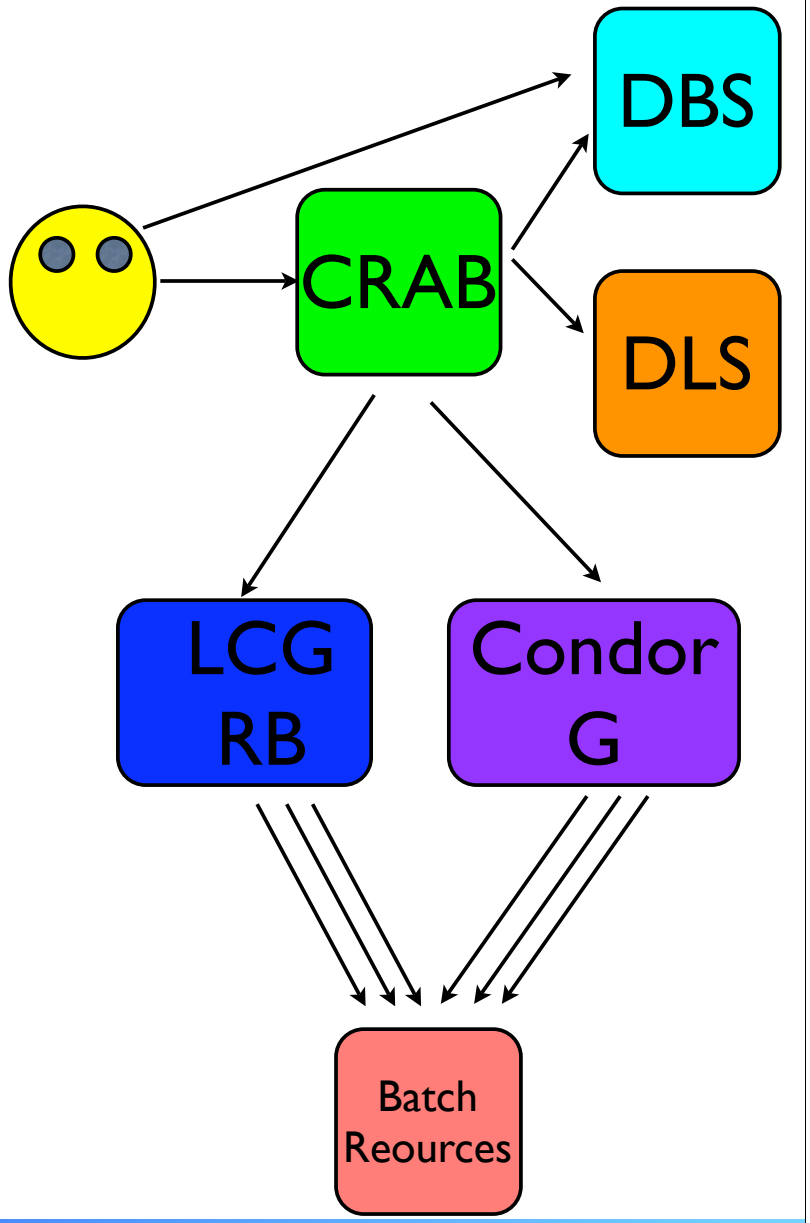
The identified dataset is defined by a number of data blocks

- ➡ Job can be sent to any site with the published set of blocks

A File list from DBS allows job splitting

Specified jobs are sent either to the LCG resource broker for the EGEE resources or Condor-G for the OSG resources

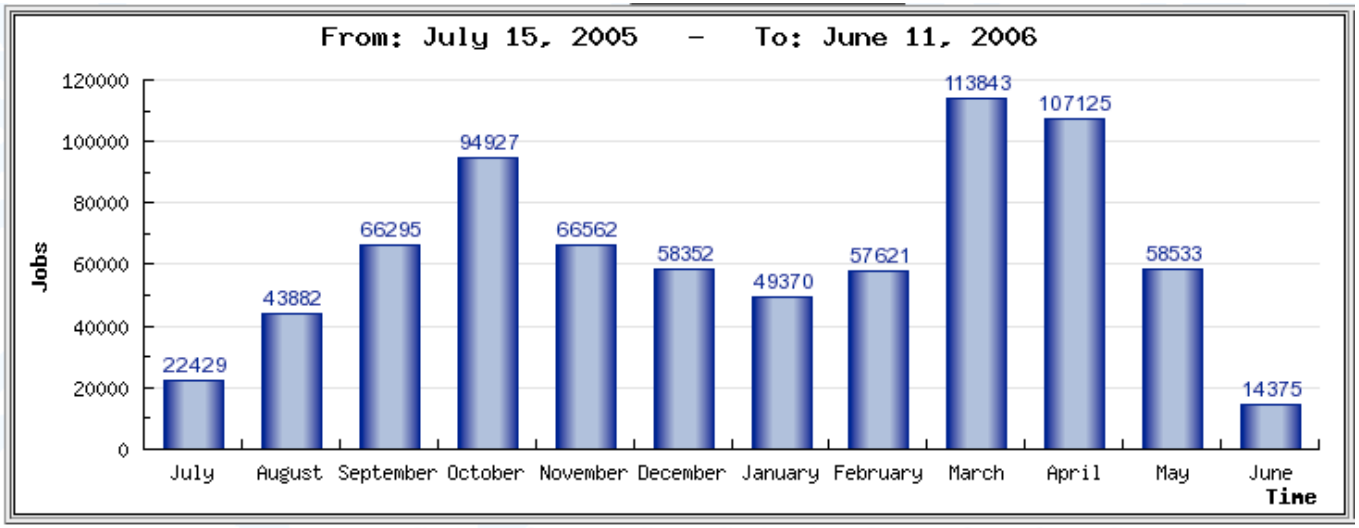
- ➡ RB has more functionality, while Condor-G is faster



Current CRAB Status

CRAB submission has reached more than 100k jobs per month

- Tends to peak before Physics TDR submissions



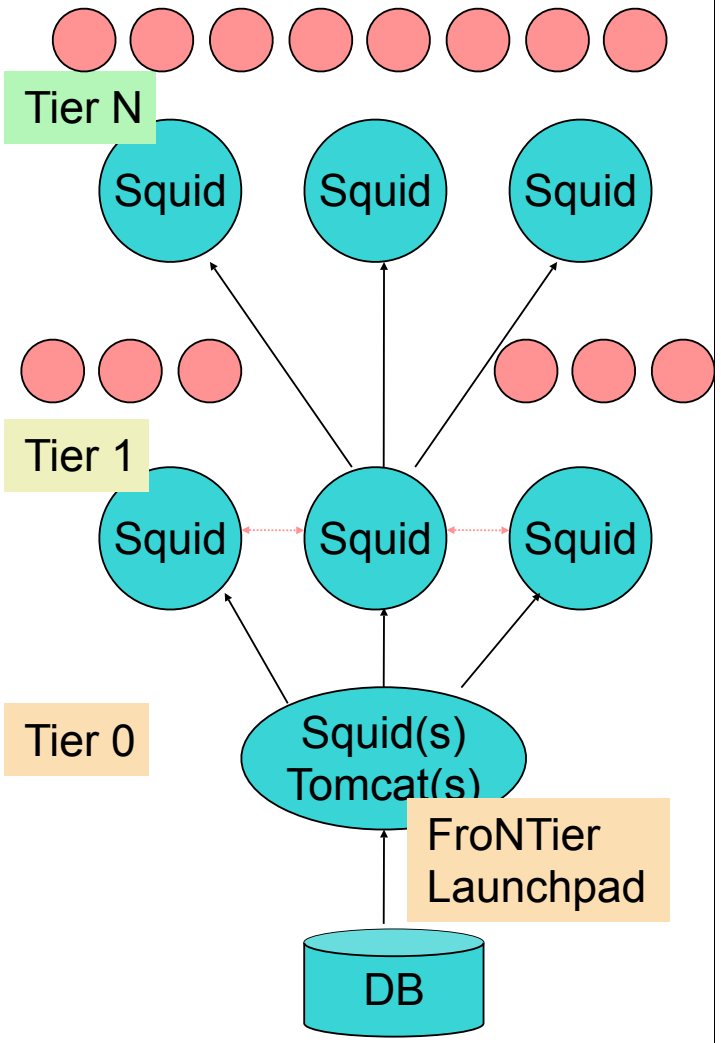
While the majority of the access has so far has been to Tier-1 centers

- A number of Tier-2 centers have hosted data samples and accepted analysis jobs
- The current demonstrated scale by users is a small fraction of the total number of jobs we expect to process

Calibration and Alignment

Successful analysis submission in a distributed environment requires access to up-to-date calibration and alignment information

- CMS is deploying and infrastructure of read-only caches
- Entire DB queries are cached
  - Very efficient if queries frequently the same
- Uses the same infrastructure for web site caching
  - Easily deployed, simplifies the DB administration and licensing at smaller sites
- Still in testing, but the initial performance numbers are promising



Future Scale

CMS expects to process approximately 100k jobs per day when the experiment is actively running in 2008

- This calculation makes a lot of assumptions about the frequency of specific kinds of data access and the number of active collaborators

We have been ramping up the number of jobs submitted through CRAB in the context of the WLCG Service Challenges

- We can currently sustain around 15k jobs per day through an infrastructure of 4 Resource brokers
  - We believe to meet the 2008 job submission goals we need to switch to the gLite RB with bulk submission
    - Tests are on-going there is a lot of validation to do.
- We can sustain a large number of submissions through Condor-G to the LCG sites, but no brokering or resource selection is available
  - Looking at gLite WMS for OSG or other resource selection techniques

Open Areas of Work

There are a number of open areas of work to enhance the current infrastructure

- Problem diagnosis and debugging
  - Job monitoring and tracking has improved in CMS, but users still have less information than a traditional batch queue
    - It's hard to tell where the failure occurred and hard to distinguish infrastructure failures from application or user problems
- We do not yet distinguish activities for prioritization
  - CMS would like to be able to specify the utilization of resources based on the experiment scientific priorities
    - Much of the underlying grid technology exists do to at least coarse divisions, but the deployment in an operationally sustainable way is difficult
- The experiment and grid infrastructures must continue to improve in reliability
  - The CMS goal for grid submission success this year is 90%.
    - Need to improve

Outlook

CMS is designing a grid analysis system where resource utilization is based on data location

- This simplifies the decision on were applications should be run
- It makes realistic expectations of the functionality of grid services

The first years of analysis are likely to be a strenuous test of the grid and experiment service infrastructure

- We believe we have choosen a computing model that is deployable

There is a large ramp in scale, performance and reliability that must be achieved before the experiment begins real data analysis

- But significant progress is being made.